云南大学学报 (自然科学版), 2002, 24 (3): 199~201 CN 53-1045/N ISSN 0258-7971
Journal of Yunnan University

统计方法求系统树*

汪 浩
(云南大学 非线性复杂系统中心, 云南 昆明 650091)

摘要: 提出比较 2 个字符序列的一个统计方法, 并将它用于细菌和古细菌的基因序列, 得出的系统树显示支持细菌和古细菌分为 2 个大的类别.

关键词: 系统树; 界; 相似性; 距离; DNA 序列; 统计方法
中图分类号: O 212.1 文献标识码: A 文章编号: 0258-7971(2002)03-0199-03

地球上的生物, 尽管形态和生活方式千差万别, 但基本生化过程有着一致性, 分子生物学发展起来以后又发现遗传密码是统一的, 这使人们相信生物有共同祖先[1]. 传统的根据形态比较作出的分类可以给出演化过程的参考, 化石研究也有助于了解物种的亲缘关系. 而分子生物学的发展, 特别是核酸与蛋白数据的大量积累, 使人们可以从分子水平追溯亲缘关系[2].

20 世纪 70 年代, Carl Woese 等在分子研究的基础上提出原核生物分为 2 个大的类别: 细菌和古细菌. 他建议把原核生物再分两界[3]. 后来分子系统树的研究已涵盖了原核生物与真核生物的大部分门类, 并揭示出建立在表型比较方法上的系统树所无法显示的物种之间的关系与进化规律[5]. 这一研究引起很大争论, 目前生物界有二界, 三界, 五界, 六界, 八界之争[6].

目前的系统树构建中, 比较特征(序列间的距离的确定) 主要用联配的打分方法. 本文提出一种比较 2 个基因序列的统计方法, 用它分析细菌和古细菌.

1 统计方法

字母表: 字母表指 1 个集合, 它的每个元称为字母. 例如 $\Delta = \{a, t, g, c\}$, 我们就称 Δ 为 4 字母的字母表.

字: 指由字母表中的字母构成的给定长度的符号串. 若字母表有 t 个元. 给定的长度为 l, 则长为 l 的字的种类数为 t^l.

窗口和窗口字: 给定长度 l, 我们就得到长度为 l 个单位的窗口. 其中每个单位允许填入 1 个字符. 当所有单位都填满后. 我们就得到 1 个字. 称为窗口字.

窗口滑动: 写出符号串 S, 给定长度 l 的窗口, 将这个窗口的第 1 个单位与 S 的首字符对齐, 第 2 个单位与 S 的第 2 个字符对齐, ……, 第 l 个单位与 S 的第 l 个字符对齐, 并把这些字符填入对应单位中, 这样就得到一个窗口字. 下一步, 将窗口的第 1 个单位 S 与的第 2 个字符对齐, 第 2 个单位与 S 的第 3 个字符对齐, ……, 第 l 个单位与 S 的第 $l+1$ 个字符对齐, 并把这些字符填入对应单位中, 得到另一个窗口字. 这时称窗口滑动了 1 个单位. 继续滑动, 直到窗口的第 l 个单位与符号串的末字符对齐, 这时得到最后一个窗口字.

特征矩阵: 给定长度为 l, 由一个字符序列用窗口滑动的办法可以得到一组窗口字. 统计各不同种类的窗口字可以得到 1 个 $2 \times t^l$ 的矩阵 W. W 的第 1 行记录这个字符序列中在窗口滑动地过程中出现的不同种窗口字, 第 2 行记录对应窗口字的数目. 称 W 为 S 的关于 l 的特征矩阵.

不同物种的特征矩阵不同. 预示可以用之分辨序列. 如果假定序列的相似性反映物种的发生学亲近程度, 则比较序列就是比较物种.

对 2 个特征矩阵进行比较, 可得它们所对应的序列之间的差异. 称为距离. 距离 $d(l, t)$ 定义为

$$d(l, t) = D(l, t)/N(l), \quad (1)$$

$N(l)$ 是归一化数, 大小为被比较的 2 个序列的窗口字数的总和. 它与 l 有关, 如 N_1 和 N_2 分别为 2

* 收稿日期: 2001-12-03
基金项目: 973(非线性科学) 的部分资助项目.
作者简介: 汪 浩(1976-), 男, 云南人, 硕士生, 主要从事符号动力学研究.

序列所含字母数, 则

$$N(l) = N_1 + N_2 - 2l + 2, \quad (2)$$

$D(l, t)$ 是 2 个序列的差异的绝对数量

$$D(l, t) = \sum_{i=1}^{t^l} | n_i^1 - n_i^2 |, \quad (3)$$

n_i^1 表示序列 1 的特征矩阵的第 i 种窗口字的数目, n_i^2 表示序列 2 的特征矩阵的同种窗口字的数目. 于是

$$d(l) = \frac{1}{N_1 + N_2 - 2l + 2} \sum_{i=1}^{t^l} | n_i^1 - n_i^2 |, \quad (4)$$

这个式子的意义是给定 l 时 2 个序列的有差异的窗口字占序列总窗口字数的比例, 显然是一个标度不变量. 可以作为衡量 2 个序列差异的特征.

2 应 用

对 1 组基因序列, 可以利用上述方法两两求出序列之间的距离, 得到距离矩阵. 然后可以作系统树. 用 EMBL[2] 中的 29 个物种的 16sRNA 由最大邻接方法[4] 作图, 物种如下: 1. Aeropyrum pernix TB1, 2. Aeropyrum pernix TB7, 3. Aeropyrum pernix, 4. M. fervidus 16S rRNA, 5. M. thermoautotrophicum (DELTA H), 6. Methanococcus jannaschii, 7. Pyrococcus horikoshii, 8. S. solfataricus, 9. T. tenax, 10. Aquifex aeolicus, 11. B. burgdorferi (297 strain), 12. Bacillus subtilis, 13. Bradyrhizobium japonicum PRY65, 14. C. pneumoniae, 15. C. ulcerans, 16. Chlamydia trachomatis, 17. Chlorobium limicola UdG 6002, 18. D. radiopugnans, 19. E. coli, 20. Flavobacterium heparinum, 21. H. aurantiacus 16S, 22. H. influenzae 16S, 23. H. pylori, 24. M. genitalium, 25. M. pneumoniae, 26. Mycobacterium turberculosis, 27. Rickettsia prowazekii, 28. Thermotoga maritima, 29. Treponema pallidum. 结果如图 1~4.

其中 1 至 9 表示的物种来自古细菌, 其余来自细菌. 又 1, 2, 3 号物种选自同一物种的不同菌株. 从图上可以看出随 l 的增长显示出细菌与古细菌明显的分别成团现象. $l = 6$ 时, 古细菌虽已经大致聚合, 但物种从共同祖先的第一次分离却不发生在细菌与古细菌之间(17, 21, 23, 2, 25 为一组, 其它为一组). $l = 7$ 以后, 我们发现物种从共同祖先的第一次分离分离都发生在细菌与古细菌之间, 说明按统计比对的结果支持细菌与古细菌应该分两界. 另外我们还注意到 Aquifex aeolicus 和 Thermotoga maritima 总和古细菌关系密切, 这有两种解释, 一是他们确实相近, 二是由于达不到最大邻接方法所要求物种数目而产生的误差所致. 因为从理论上说只有用全部的物种才能得到精确结果.

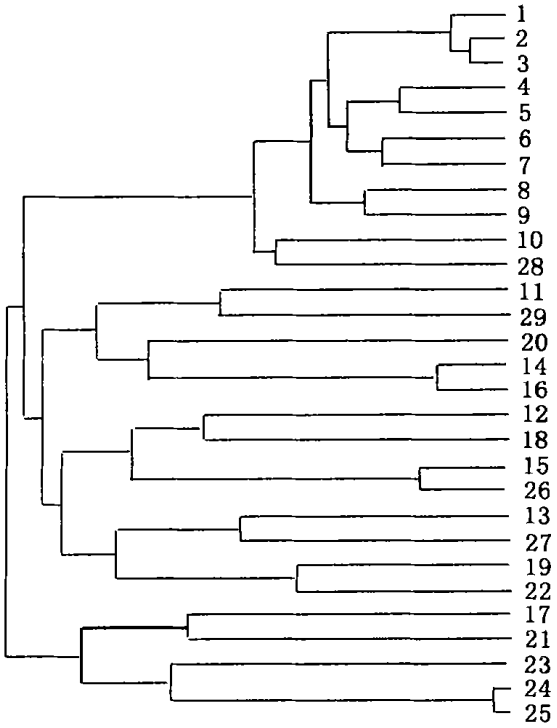

图 1 $l = 6$ 的系统树

Fig. 1 Phylogenetic tree with $l = 6$

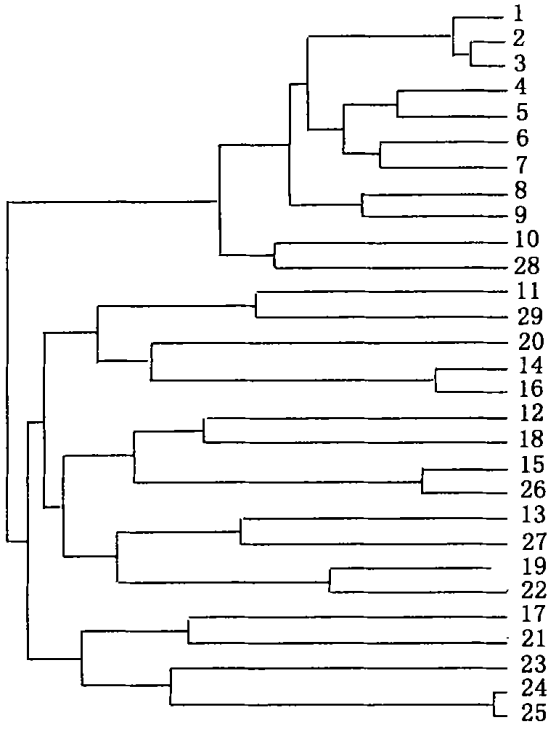

图 2 $l = 7$ 的系统树

Fig. 2 Phylogenetic tree with $l = 7$

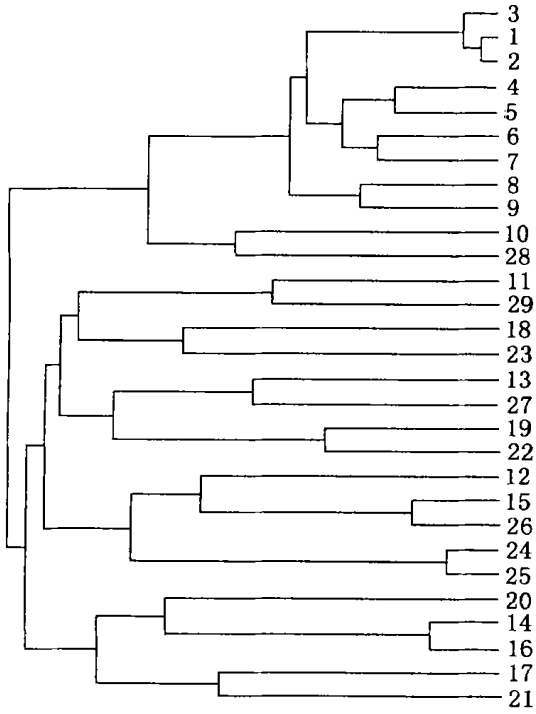

图 **3**　l **= 8** 的系统树

Fig. 3 Phylogenetic tree with l = 8

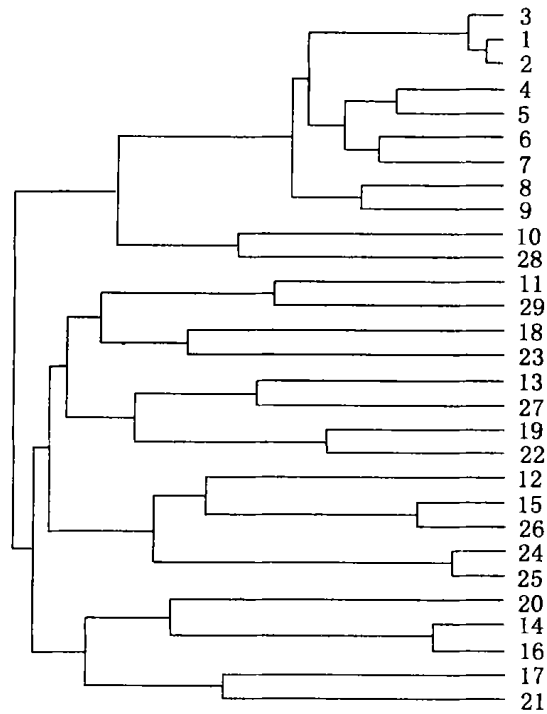

图 **4**　l **= 9** 的系统树

Fig. 4 Phylogenetic tree with l = 9

3　讨　论

从上面的分析可以看出这个方法只有在 l 被适当选择时才有效, 过小和过大的 l 都将导致无法分辨序列. 关于 l 的取值范围的讨论是下一步的工作.

参考文献:

[1] 郝柏林, 张淑誉. 生物信息学手册[M]. 上海: 上海科学技术出版社, 2000.

[2] http://www.ebi.ac.uk/embl/, 2001-12-18.

[3] WOESE C R, KANDLER O, WHEELIS M L. Towards a Natural System of organisms: proposal for the domains archaea, bacteria and eucarya. [J] Proc Natl Acad Sci USA, 1990, 87: 4 576—4 579.

[4] NARUYA S, MASATOSHI N. The neighbor joining method: a new method for reconstructing phylogenetic trees[J]. Mol Boil Evol, 1987, 4: 406—425.

[5] FOX G E, STACKEBRANDT E, HESPELL R B, et al. The phylogeny of prokaryotes[J]. Science, 1980, 209: 457—463.

[6] CAMPBELL N A. Biology 4th ed[M]. San Francisco: Benjamin/Cummings, 1996.

Phylogenetic analysis with a statistical method

WANG Hao

(Center for Nonlinear Complex Systems, Department of Physics, Yunnan University, Kunming 650091, China)

Abstract: A new statistical method to analyze symbolic sequences is proposed. The DNA sequences of prokaryotes are analyzed. The phylogenetic analysis shows that Bacteria and Archaea are separated into two groups, which supports Carl Woese's "there kingdoms" theory.

Key words: phylogenetic tree; kingdoms; similarity; distance; DNA sequences; statistical method